Multinomial logistic regression report

Team: Who's there

Introduction

The aim of the project was to be able to find a model that best predict the relationship between: * the number of people counted with the survey for a given class at a particular hour * Wi-fi log counted in that room at that hour

This will allow to see whether Wi-Fi log is a good predictor for estimating occupancy in a classroom.

Given the fact that the target feature was a percentage we tried to run a logistic regression to predict when the room was more likely to be empty or occupied.

Below we describes step by step all the analysis performed.

DATABASE CONNECTION AND DATASET

First of all, we set up the connection to the database, using the following code:

```
connection <- dbConnect(MySQL(),user="root", password="",dbname="who_there_db", host="localhost")
```

Then we made a query to the database, in order to get all the groundth truth data collected in room B.002, B.004 and B.006 from 9 to 17 and the correspondent Wi-Fi Log measured in that time frame and rooms.

The dataset created had in total 216 rows and it will allow us to explore if Wi-Fi log can estimate precisely if a room is empty or occupied during a certain hour.

As a **target features** for the multinomial logistic regression we decided to create a categorical features with 4 levels using the percentage of the room full. In particular, the levels were: * Low occupancy with an occupancy between 0 and 25%; * Mid Low occupancy, which ranges from 25-50% of occupancy; * Mid High occupancy, which contain obeservation between 50-75%; * High Occupancy with an occupancy range between 75-100%.

As response variables or feature we considered Wi-Fi logs, which were summarised either as average of the logs counted for each room and for each hour or as maximum of the logs measured for each room and for each hour.

Together with the Wi-Fi log, we included in the data set the following features:

- **Date**, which we did not use in this analysis, because they just cover 2 weeks of Novemeber, but for future analyses they can be used to group observations by seasons or semesters or to finds seasonal trends for time series analyses.
- **Time**, which will be explored either as continous variable and as categorical to explore if the time of the day can have an affect on the Wi-Fi log. To do so we, bin the time in 4 ranges: early morning (9-11), late morning (11-13), early afternoon (13-15) and late afternoon (15-17). This will allow us to see if the Wi-Fi log accuracy was changing during the day. For example, it is more likely that all the electronic devices are fully powered early in the morning and consequently the Wi-fi log data can be more accurate or overestimating the occupancy of the room (i.e. more than one device per person). On the contrary in the afternoon, the devices may be more likely to be out of battery and it is possible that there are less devices in the room.

- **Module**, which we are not going to include in the analysis because the majority of the module present are for computer science, but for future analyses it will be possible to explore if the Wi-fi log accuracy in predicting the occupancy change acroos the courses. Science course or computer science course will more likely to use electronic devices during lecture than art students.
- **Course level**, which can indicate us whether electronic devices will be less used during different course level. For example, first and second level courses can be less practical and therefore laptop are not needed and that can decrease the number of devices connected. On the other hand, undergraduate might be more distracted during lecture and look at their phones during lectures. This will result in an increase of connection in that hour.
- **Tutorial**, which can affect the number of logged people. First of all, because tutorial divided the room in 2 and therefore there will be measured less people than expected.
- **Double_module**, categorical variable indicating wether in the class there are more than one module, increasing the number of people expected in the room.
- **Double_module**, categorical variable indicating wether in the class went ahead to check for false positive.

The resulting data set is printed below:

```
head(AnalysisTable)
```

```
##   Room       Date Time    Module Course_Level Tutorial Double_module
## 1    1 2015-11-03    9         0            0        0             0
## 2    1 2015-11-04    9 COMP30190            3        0             0
## 3    1 2015-11-05    9         0            0        0             0
## 4    1 2015-11-06    9 COMP30220            3        0             0
## 5    1 2015-11-09    9 COMP30190            3        0             0
## 6    1 2015-11-10    9         0            0        0             0
##   Class_went_ahead Capacity Percentage_room_full Wifi_Average_logs
## 1                1       90                 0.00            4.7500
## 2                1       90                 0.25           13.4545
## 3                1       90                 0.00            6.8333
## 4                1       90                 0.00            2.4167
## 5                1       90                 0.25           14.7273
## 6                1       90                 0.00            2.2727
##   Wifi_Max_logs Binned_Occupancy   Factor_Time
## 1            21              Low Early Morning
## 2            15          Mid_Low Early Morning
## 3            29              Low Early Morning
## 4             3              Low Early Morning
## 5            18          Mid_Low Early Morning
## 6            14              Low Early Morning
```

DATA QUALITY REPORT

Before running any analyses, we carried out the data quality report to check for any issue related to the variable (e.g. outlier, skewed distribution, NaN values) and solutions we will implement to solve them.

Initially set all the categorical variables as factors and then we printed the descriptive statistic for all the features.

```
summary(AnalysisTable)
```

```
##       Room       Date                Time          Module
##  Min.   :1   Length:216         Min.   : 9.00   Length:216
##  1st Qu.:1   Class :character   1st Qu.:10.75   Class :character
##  Median :2   Mode  :character   Median :12.50   Mode  :character
##  Mean   :2                      Mean   :12.50
##  3rd Qu.:3                      3rd Qu.:14.25
##  Max.   :3                      Max.   :16.00
##  Course_Level    Tutorial       Double_module Class_went_ahead
##  0:59         Min.   :0.00000   0:210         0: 22
##  1:14         1st Qu.:0.00000   1:  6         1:194
##  2:23         Median :0.00000
##  3:76         Mean   :0.02778
##  4:40         3rd Qu.:0.00000
##  5: 4         Max.   :1.00000
##     Capacity     Percentage_room_full Wifi_Average_logs Wifi_Max_logs
##  Min.   : 90.0   Min.   :0.00         Min.   :  0.00    Min.   :  0.00
##  1st Qu.: 90.0   1st Qu.:0.00         1st Qu.: 11.61    1st Qu.: 18.75
##  Median : 90.0   Median :0.25         Median : 23.67    Median : 32.50
##  Mean   :113.3   Mean   :0.25         Mean   : 30.33    Mean   : 40.01
##  3rd Qu.:160.0   3rd Qu.:0.25         3rd Qu.: 43.88    3rd Qu.: 55.25
##  Max.   :160.0   Max.   :1.00         Max.   :192.92    Max.   :230.00
##  Binned_Occupancy          Factor_Time
##  Low     :67      Early Morning  :54
##  Mid_Low :96      Late Morning   :54
##  Mid_High:40      Early Afternoon:54
##  High    :13      Late Afternoon :54
##
##
```

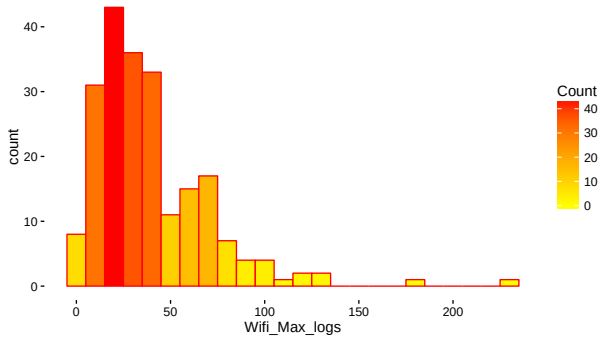
From this we could see that NaN values were not present in the data set. We could notice that the observations for the features Tutorials and Double_model were not even distributed across the 2 levels of the variables. In fact, only 6 observations were present for tutorial class and for double module class. Therefore, we decided to discard both the features, because they will be not informative for the analysis. Similarly for the feature class_went_ahead the majority of the lectures did occur and we decided to discard it. Furthermore for the variables Wifi_Average_logs and Wifi_Max_logs, it seems that there are few outliers, since the median is lower than the mean and the max values are far higher than the mean values. We will going to explore this issues with histogram and boxplots.

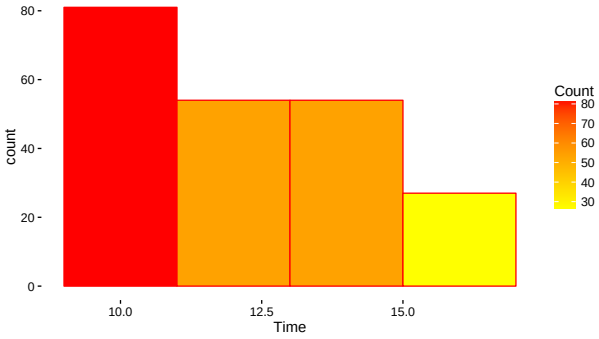
Exploratory graphs

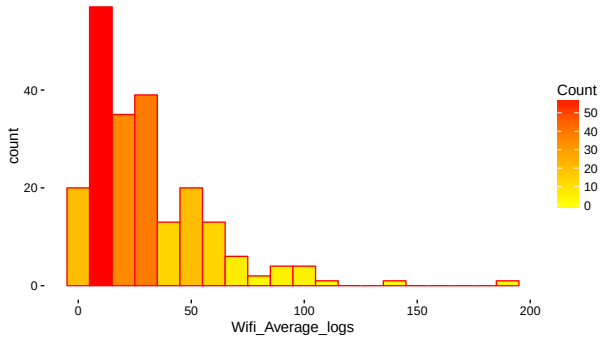
For exploring possible issues related with the continuous variables we plotted histograms and boxplots.

Histograms

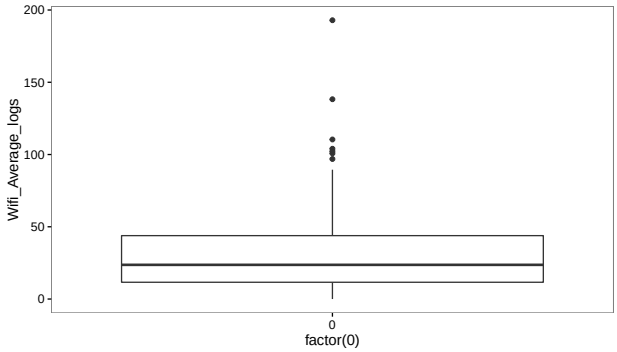
```
## Loading required package: grid
```

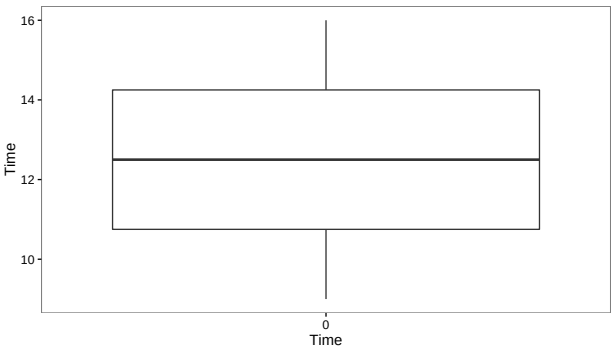


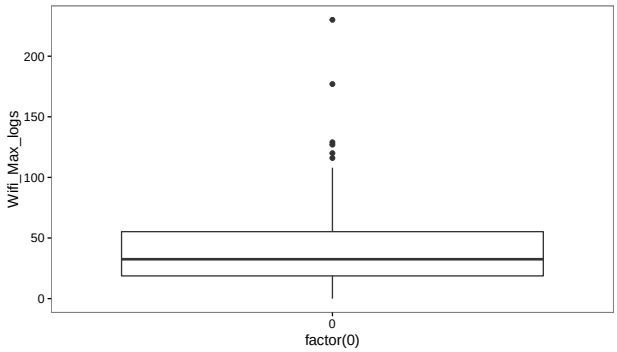




Form the histograms we could see that the distribution of the feature Wifi Maximum_client (i.e. the Maximum number of devices logged in one hour lecture) was skewed to the left, indicating that the in the majority of the lecture were counted no more than 40 people. Furthermore, we could see that there are potential outliers (values > 150). Similar pattern was observed for the feature Wifi_Average_logs. Feature times had as well a skewed distribution, suggesting that the majority of the lectures were concentrating during the early morning and they were decreasing towards the afternoon. ### Box plots.

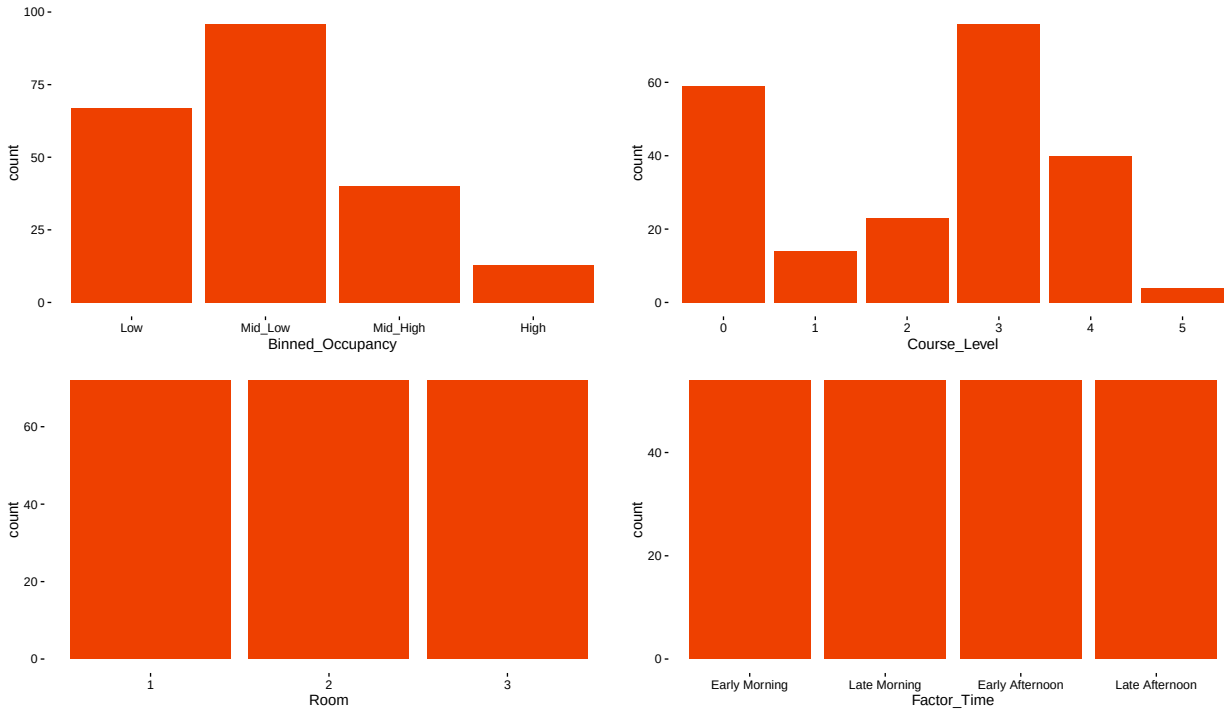






From the boxplots, all the trends observed in the histograms were confirmed.

For categorical variables we plot bar plot graphs.



From the barplots, we could see that observations were equally distributed across all the levels of the feature Room and Factor Time. On the contrary, there were more observations for non lectures and level 3 courses and classes was observed mostly occupied. No issues were detected for those features.

Summary.

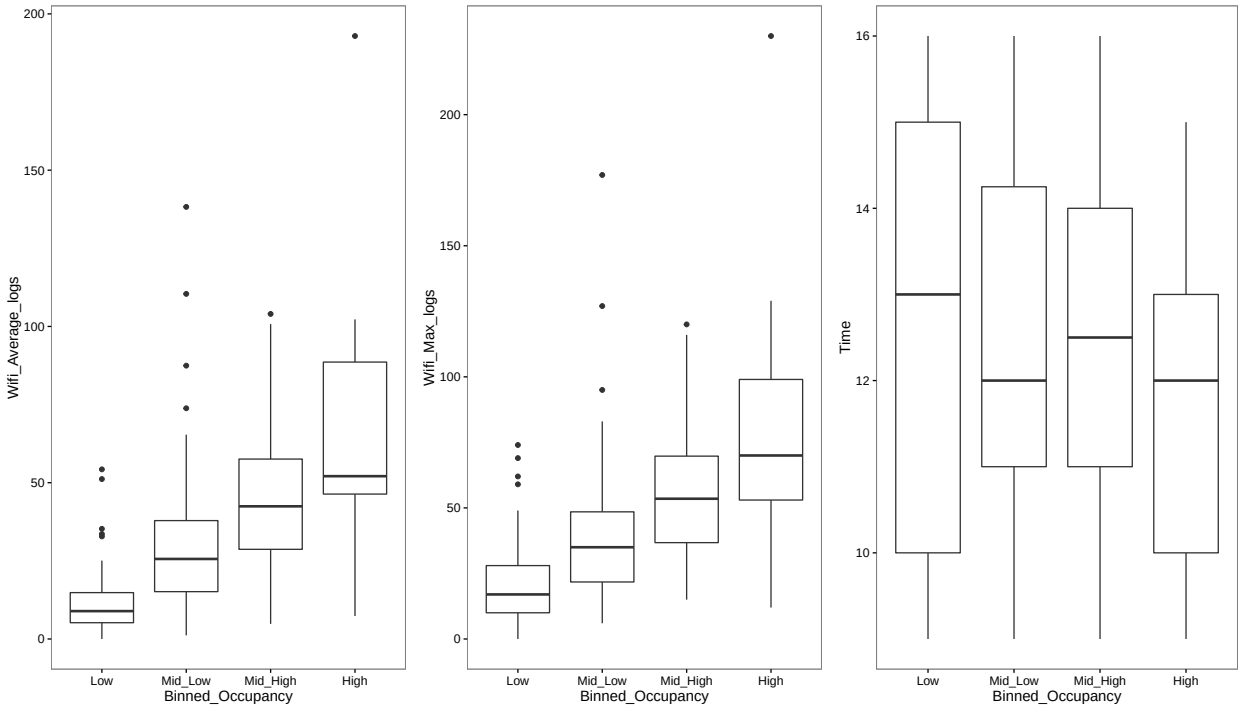
Features	Issues	Planned Solution
Room	None	None
Time	Distibution skewed to the left	To solve during analysis
Factor Time	None	None
Course level	None	None
Tutorial	Uneven representation of the level	Discarded from the analysis
Double Module	Uneven representation of the level	Discarded from the analysis
Class went ahead	Uneven representation of the level	Discarded from the analysis
Wifi Average clients	Distibution skewed to the left & outliers	To solve during analysis
Wifi Maximum clients	Distibution skewed to the left & outliers	To solve during analysis
Binary Occupancy	None	None

FEATURES AFFECTING THE TARGET FEATURE

The next step of the analysis was to see which feature really affect the target feature for deciding which features we would include into the model.

For the continuous features we used box plots, while for categorical we use bar plots.

Box plots for categorical variables.



From the boxplots we could see that when the room was indicated as empty either the average and the maximum Wi-Fi logs were close to zero and they were increasing across the level of the binned occupancy, showing that either average Binary_Occupancy and maximum wifi counted clients are a good predictor of the binary occupancy and we are going to run 2 models: one for exploring the relationship between Binary_Occupancy and Average counted clients and another for Survey counted clients and Maximum counted clients. The difference between the levels will probably be higher without the outliers and we might run the model without them to see if the accuracy will increase. The occupacy seems to not change with the time, so were not going to explore it.

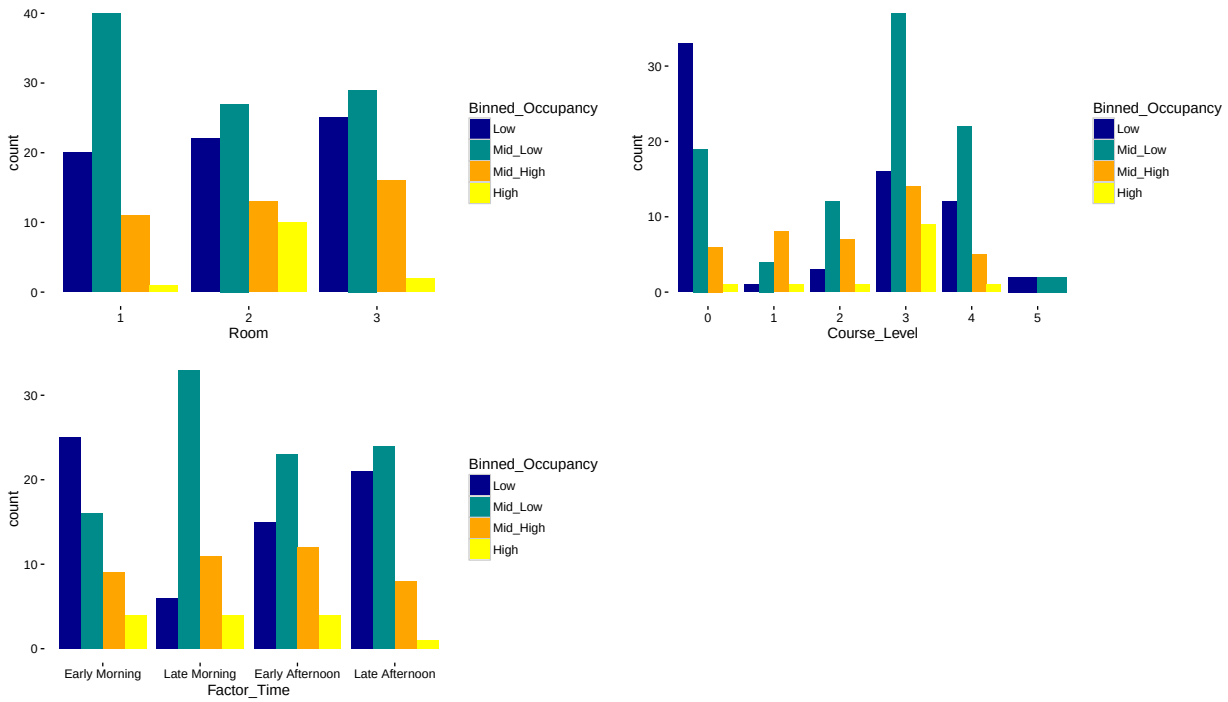
Barplots

```
barpair1 <-ggplot(AnalysisTable, aes(x = Room, fill = Binned_Occupancy)) + geom_bar(position = "dodge")+

barpair2 <-ggplot(AnalysisTable, aes(x = Factor_Time, fill =Binned_Occupancy)) + geom_bar(position = "d

barpair3 <-ggplot(AnalysisTable, aes(x = Course_Level, fill = Binned_Occupancy)) + geom_bar(position =

multiplot(barpair1, barpair2, barpair3, cols=2)
```



All the ranges of occupancy were equally represented in all the three rooms, indicating that room was not affecting the ocuupancy. The occupancy level tend to change across the levels of course levels. Level1 and level2 courses had an higher Mid-low and Mid-high occupancy, while all the other levels had a higher low and mid-low occupancy. This was expected for level 0, where we expect no lecture is running. Therefore, we are considering all those features in the model.

Interacting effect on the target feature

For selecting the features that together with the max logs or with the average logs best predict the ground truth data we will do model selection using AIC values.
For doing so, we decided to considered the models that better suited our hypothesis and they are the following:

- *Binned occupancy ~ 1*, the null model;
- *Binned occupancy ~ Average Wifi occupancy*, for testing whether average Wi-Fi counting logs were accurately predicting the occupancy of the room;
- *Binned occupancy ~ Average Wifi occupancy + Room*, for testing whether average Wi-Fi counting logs and the room were were accurately predicting the occupancy of the room;
- *Binned occupancy ~ Average Wifi occupancy + Time*, for testing whether average Wi-Fi counting logs and the time of the day were accurately predicting the occupancy of the room;
- *Binned occupancy ~ Average Wifi occupancy + Course_Level*, for testing whether average Wi-Fi counting logs and course levels were accurately predicting the occupancy of the room;
- *Binned occupancy ~ Average Wifi occupancy + Room + Time*, for testing whether average Wi-Fi counting logs, room type and the time of the day were accurately predicting the occupancy of the room;
- *Binned occupancy ~ Average Wifi occupancy + Room + Course_Level*, for testing whether average Wi-Fi counting logs, room type and course levels were accurately predicting the occupancy of the room;

- *Binned occupancy ~ Average Wifi occupancy + Time + Course_Level*, for testing whether average Wi-Fi counting logs, the time of the day and the course levels were accurately predicting the occupancy of the room;
- *Binned occupancy ~ Average Wifi occupancy + Room + Time + Course_Level*, for testing whether average Wi-Fi counting logs, rooms, the time of the day and the course levels were accurately predicting the occupancy of the room;

The same models were run also with the occupancy estimated with the maximum number of logs measured in that hour. All these models were run using the k-fold cross. K-fold validation was preferred over the validation set approach and the Leave Out Cross Validation (LOOCV), because it is more robust and more accurate in estimating the test error. The Validation set approach tends to give an over estimate of the test error and the test error is dependent on the observations included randomly in the test set. Furthermore, the LOOCV tends to provide a test error with a high variance, because the folds used to calculating it are correlated among each other. In particular in this analysis we are going to perform a 10-fold cross validation, which is pretty standard.

The 10 fold cross validation was carried out manually using the package *carret*, as showed in the following tutorial: (http://amunategui.github.io/multinomial-neuralnetworks-walkthrough). For each model we estracted the overall accuracy and the AIC and we picked as best model the model with the lowest MSE and AIC. We are aware that for a better comparison among models is better to use AICc, but the package used only provided AIC estimates.

Results

CASE1:

All the models ran with the response variable Wifi average logs are summarised in the following table showing their accuracy and AIC.

Models	MSE
Binned_occupancy ~ 1	0.465
Binned_occupancy ~ Wifi_Average_logs	0.56
Binned_occupancy ~ Wifi_Average_logs + Room	0.575
Binned_occupancy ~ Wifi_Average_logs + Factor_Time	0.535
Binned_occupancy ~ Wifi_Average_logs + Course_Level	0.535
Binned_occupancy ~ Wifi_Average_logs + Room + Factor_Time	0.545
Binned_occupancy ~ Wifi_Average_logs + Room + Course_Level	0.575
Binned_occupancy ~ Wifi_Average_logs + Factor_Time + Course_Level	0.52
Binned_occupancy ~ Wifi_Average_logs + Room + Factor_Time + Course_Level	0.555

The best model in term of accuracy and AIC was Binned_occupancy ~ Wifi_Average_logs + Room with accuracy equal to 0.58 and AIC equal to 364.

CASE2: MAX MODELS

Models	MSE
Binned_occupancy ~ 1	0.465
Binned_occupancy ~ Wifi_Max_logs	0.51
Binned_occupancy ~ Wifi_Max_logs + Room	0.525
Binned_occupancy ~ Wifi_Max_logs + Factor_Time	0.5
Binned_occupancy ~ Wifi_Max_logs + Course_Level	0.5
Binned_occupancy ~ Wifi_Max_logs + Room + Factor_Time	0.525
Binned_occupancy ~ Wifi_Max_logs + Room + Course_Level	0.535
Binned_occupancy ~ Wifi_Max_logs + Factor_Time + Course_Level	0.505
Binned_occupancy ~ Wifi_Max_logs + Room + Factor_Time + Course_Level	0.525

The best model in term of accuracy and AIC was Binned_occupancy ~ Wifi_Max_logs + Room with an accuracy of 0.55 and an AIC of 385. However the accuracy and the AIC of the Binned_occupancy ~ Wifi_Average_logs + Room model was better and we keep it as best model and we will run it on the whole dataset.

final.avg.room <- multinom(Binned_Occupancy ~ Wifi_Average_logs+Room, data=AnalysisTable,maxit=1000)

ORDINAL REGRESSION

Models	MSE
Binned_occupancy ~ 1	0.465
Binned_occupancy ~ Wifi_Average_logs	0.52
Binned_occupancy ~ Wifi_Average_logs + Room	0.56
Binned_occupancy ~ Wifi_Average_logs + Factor_Time	0.48
Binned_occupancy ~ Wifi_Average_logs + Course_Level	0.48
Binned_occupancy ~ Wifi_Average_logs + Room + Factor_Time	0.53
Binned_occupancy ~ Wifi_Average_logs + Room + Course_Level	0.55
Binned_occupancy ~ Wifi_Average_logs + Factor_Time + Course_Level	0.495
Binned_occupancy ~ Wifi_Average_logs + Room + Factor_Time + Course_Level	0.5

Binned_occupancy ~ Wifi_Average_logs + Room was the best model in term of AIC (377) and accuracy (0.575). Howevre, the ordinal regression did not improve the accuracy of the multinomial logistic regression. Similar pattern were expected from running the same analyses, but with Wifi_Max_logs as dependent variable.

CASE2: MAX MODELS

Models	Accuracy and AIC
Binned_occupancy ~ 1	0.465
Binned_occupancy ~ Wifi_Max_logs	0.51
Binned_occupancy ~ Wifi_Max_logs + Room	0.535
Binned_occupancy ~ Wifi_Max_logs + Factor_Time	0.49
Binned_occupancy ~ Wifi_Max_logs + Course_Level	0.49
Binned_occupancy ~ Wifi_Max_logs + Room + Factor_Time	0.525
Binned_occupancy ~ Wifi_Max_logs + Room + Course_Level	0.535
Binned_occupancy ~ Wifi_Max_logs + Factor_Time + Course_Level	0.52

Models	Accuracy and AIC
Binned_occupancy ~ Wifi_Max_logs + Room + Factor_Time + Course_Level	0.495

The best model in term of AIC (391) and accuracy (0.57) was Binned_Occupancy ~ Wifi_Max_logs+Room. The occuracy of this model was not better than the accuracy of the ordinal model with Wifi_Average_logs as dependent variable.

Therefore, we are going to considered as best model: Binned_Occupancy ~ Wifi_Average_logs+Room run with a multinomial logistic regression, suggesting that the occupancy of the room was accurately predicted by the average of the Wi-Fi logs measured across the hour. However, the accuracy of the estimate was affected by the room.